Cell Systems, Volume *3*

Supplemental Information

Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements

Sahand Hormoz, Zakary S. Singer, James M. Linton, Yaron E. Antebi, Boris I. Shraiman, and Michael B. Elowitz

Figure S1

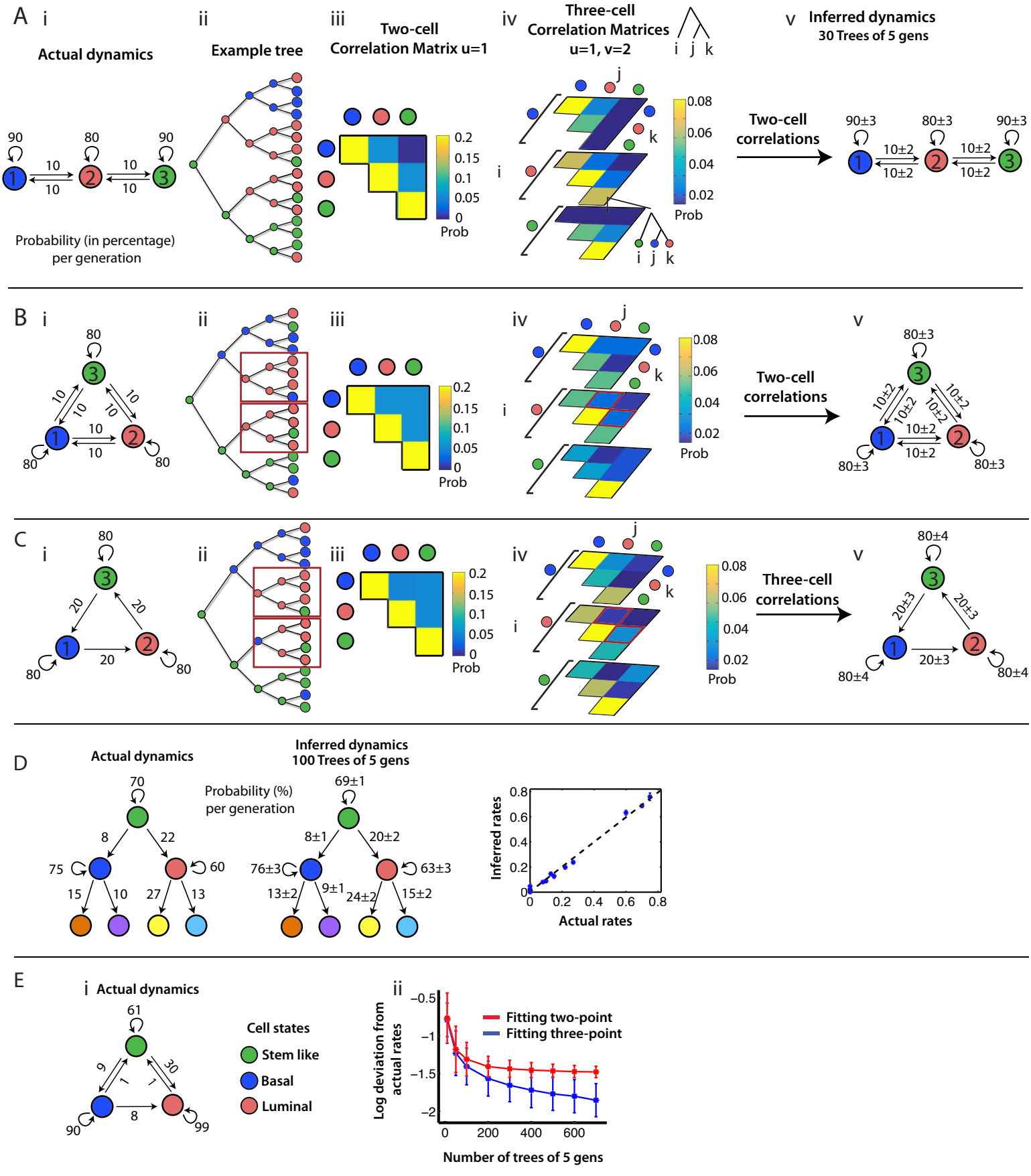

A
i
Actual dynamics
90
80
90
10
10
10
10
Probability (in percentage) per generation
ii
Example tree
iii
Two-cell Correlation Matrix u=1
0.2
0.15
0.1
0.05
0
Prob
iv
Three-cell Correlation Matrices u=1, v=2
i j k
0.08
0.06
0.04
0.02
Prob
v
Inferred dynamics
30 Trees of 5 gens
Two-cell correlations
90±3
80±3
90±3
10±2
B
80
10±2
80±3
C
20
80±4
20±3
Three-cell correlations
D
Actual dynamics
Inferred dynamics
100 Trees of 5 gens
Probability (%) per generation
70
8
22
75
60
15
10
27
13
69±1
8±1
20±2
76±3
63±3
13±2
9±1
24±2
15±2
Inferred rates
Actual rates
0.8
0.6
0.4
0.2
0
E
i Actual dynamics
61
9
1
30
1
8
90
99
Cell states
Stem like
Basal
Luminal
ii
Log deviation from actual rates
−0.5
−1
−1.5
−2
Fitting two-point
Fitting three-point
0
200
400
600
Number of trees of 5 gens

Figure S2

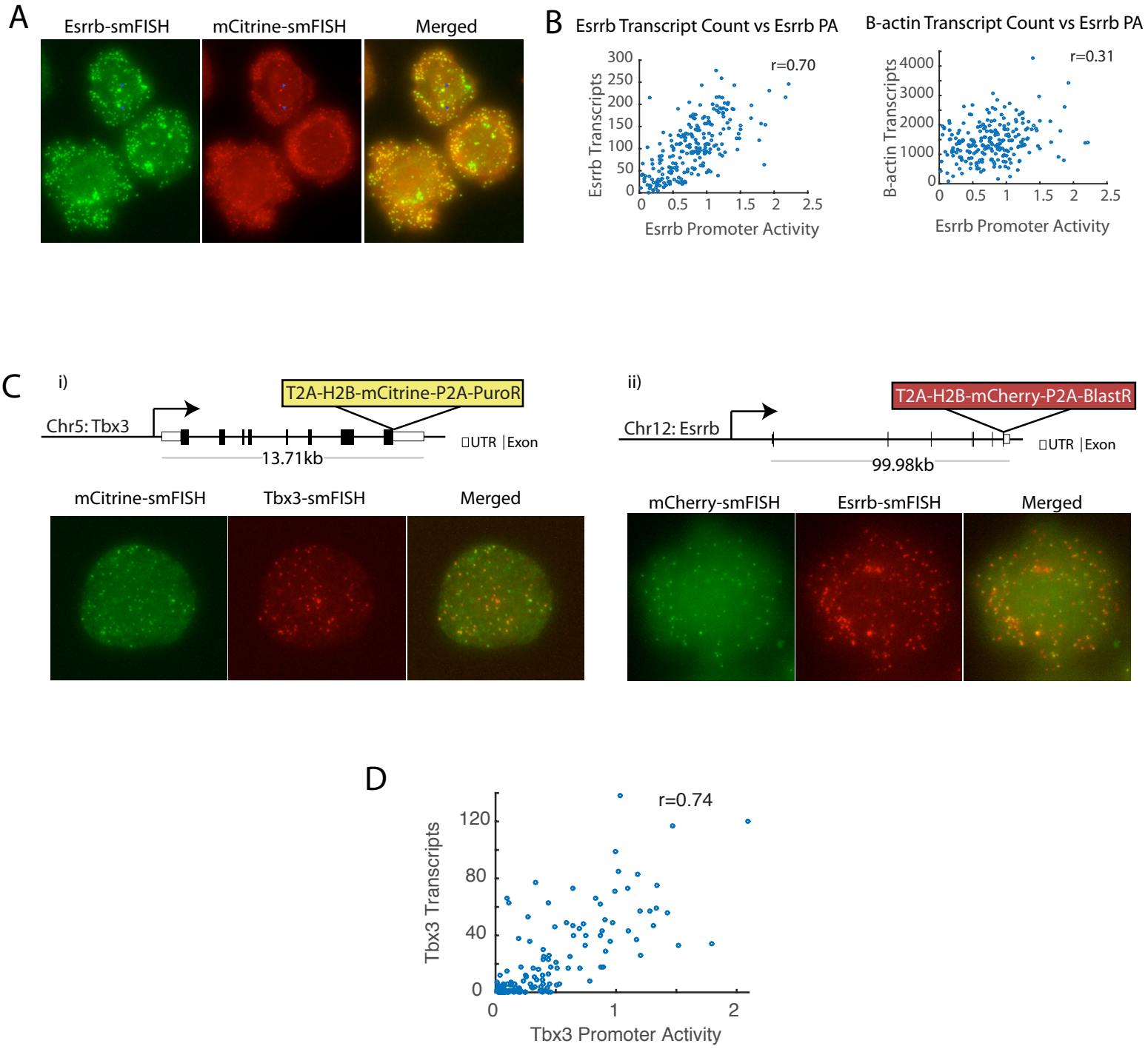

A
Esrrb-smFISH
mCitrine-smFISH
Merged
B
Esrrb Transcript Count vs Esrrb PA
r=0.70
Esrrb Transcripts
Esrrb Promoter Activity
B-actin Transcript Count vs Esrrb PA
r=0.31
B-actin Transcripts
Esrrb Promoter Activity
C
i)
T2A-H2B-mCitrine-P2A-PuroR
Chr5: Tbx3
13.71kb
UTR |Exon
mCitrine-smFISH
Tbx3-smFISH
Merged
ii)
T2A-H2B-mCherry-P2A-BlastR
Chr12: Esrrb
99.98kb
UTR |Exon
mCherry-smFISH
Esrrb-smFISH
Merged
D
r=0.74
Tbx3 Transcripts
Tbx3 Promoter Activity

Figure S3

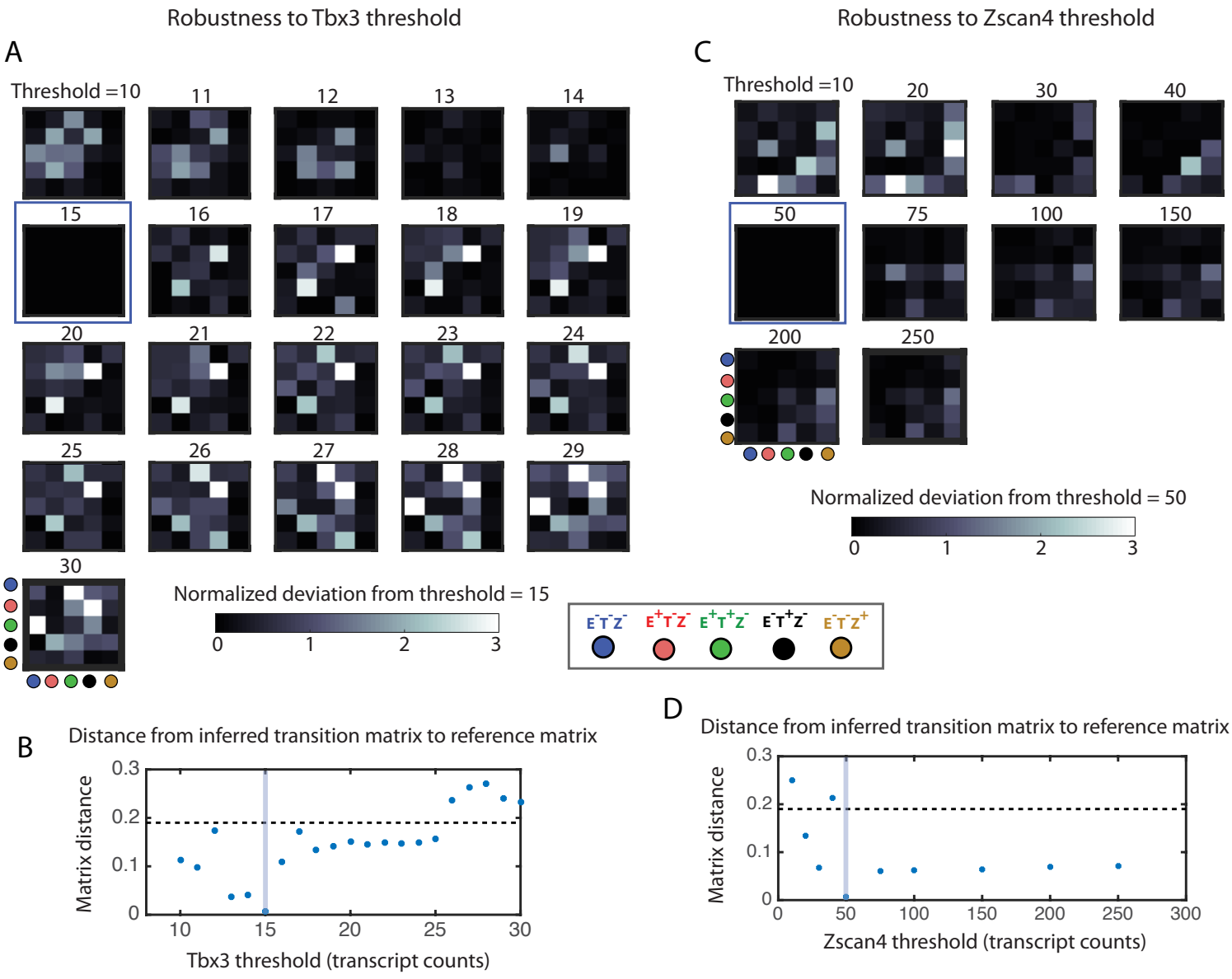

Robustness to Tbx3 threshold
A
Threshold =10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
Normalized deviation from threshold = 15
0
1
2
3
E-T-Z-
E+T-Z-
E+T+Z-
E-T+Z-
E-T-Z+
Robustness to Zscan4 threshold
C
Threshold =10
20
30
40
50
75
100
150
200
250
Normalized deviation from threshold = 50
0
1
2
3
B
Distance from inferred transition matrix to reference matrix
Matrix distance
0.3
0.2
0.1
0
10
15
20
25
30
Tbx3 threshold (transcript counts)
D
Distance from inferred transition matrix to reference matrix
Matrix distance
0.3
0.2
0.1
0
0
50
100
150
200
250
300
Zscan4 threshold (transcript counts)

Figure S4

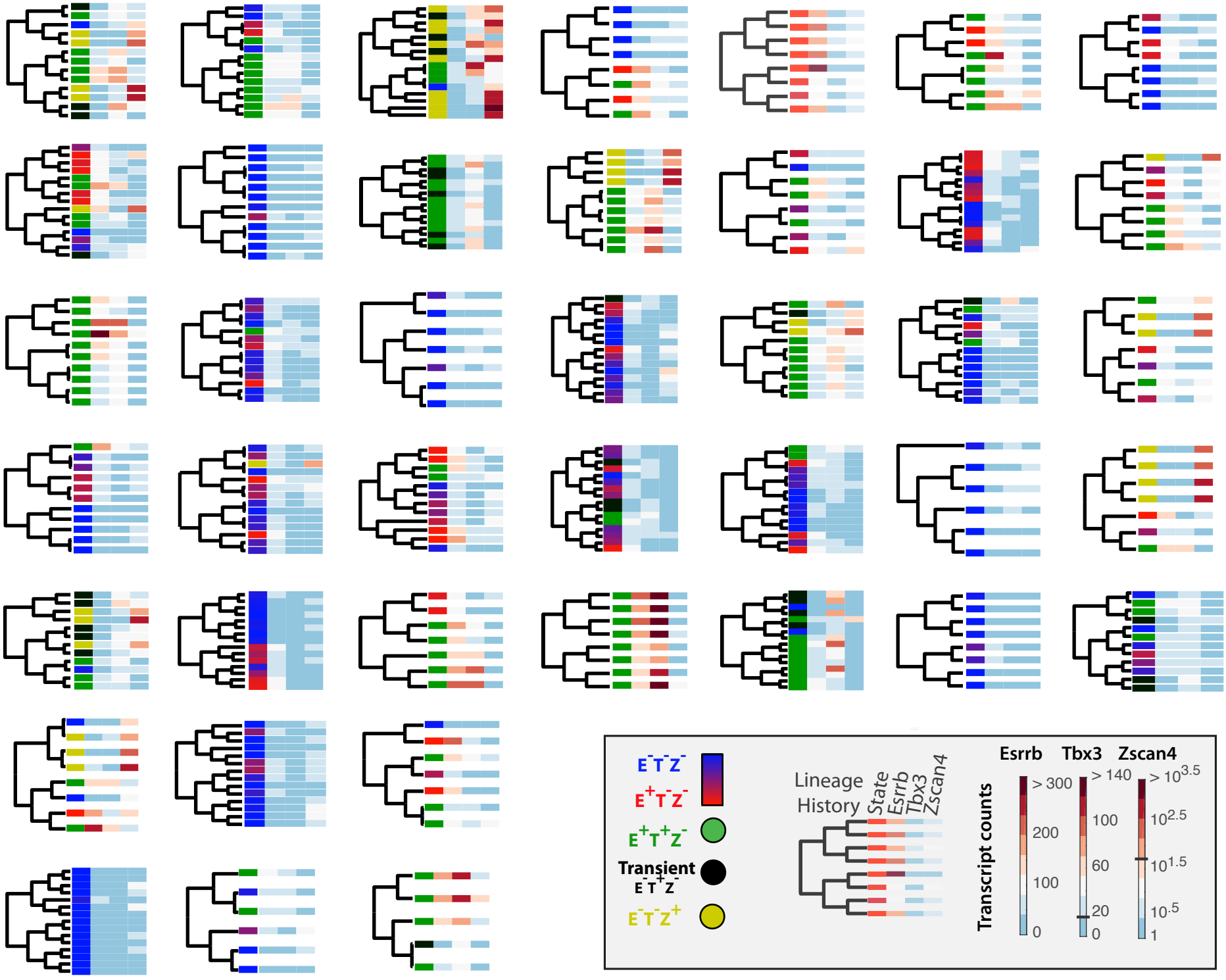

E⁻T⁻Z⁻
E⁺T⁻Z⁻
E⁺T⁺Z⁻
Transient
E⁻T⁺Z⁻
E⁻T⁻Z⁺
Lineage History
State
Esrrb
Tbx3
Zscan4
Esrrb
Tbx3
Zscan4
Transcript counts
> 300
200
100
0
> 140
100
60
20
0
> 10^3.5
10^2.5
10^1.5
10^.5
1

Figure S5

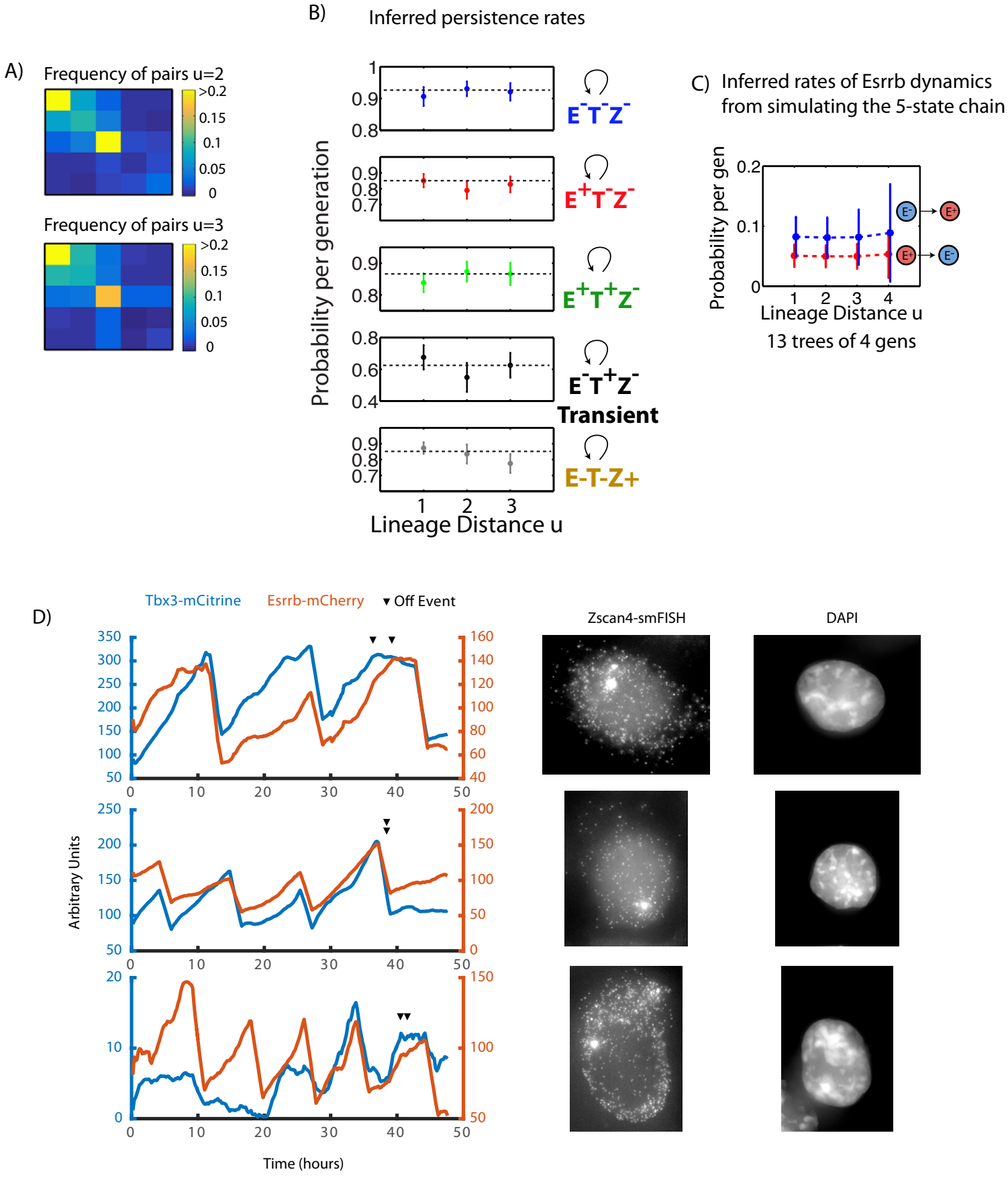

A)
Frequency of pairs u=2
>0.2
0.15
0.1
0.05
0
Frequency of pairs u=3
>0.2
0.15
0.1
0.05
0
B)
Inferred persistence rates
Probability per generation
E-T-Z-
E+T-Z-
E+T+Z-
E-T+Z-
Transient
E-T-Z+
Lineage Distance u
C)
Inferred rates of Esrrb dynamics
from simulating the 5-state chain
Probability per gen
Lineage Distance u
13 trees of 4 gens
D)
Tbx3-mCitrine
Esrrb-mCherry
Off Event
Arbitrary Units
Time (hours)
Zscan4-smFISH
DAPI

Figure S6

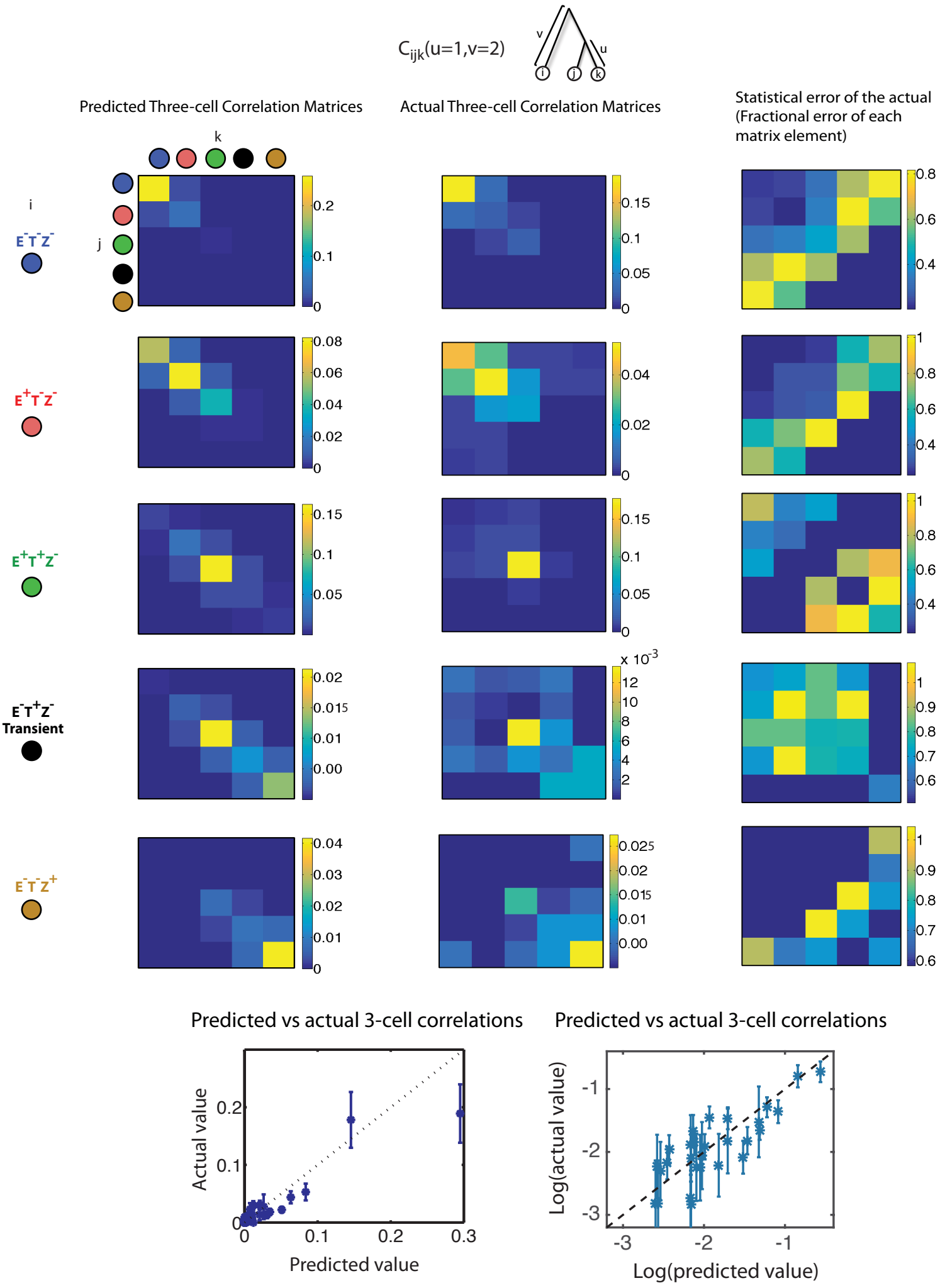
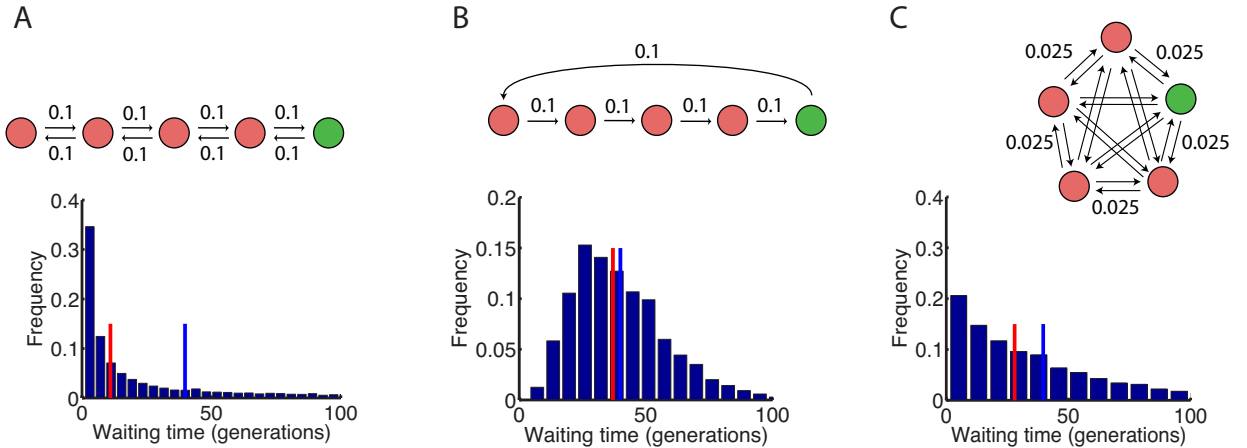

$C_{ijk}(u=1,v=2)$
v
u
i
j
k
Predicted Three-cell Correlation Matrices
Actual Three-cell Correlation Matrices
Statistical error of the actual (Fractional error of each matrix element)
$E^-T^-Z^-$
$E^+T^-Z^-$
$E^+T^+Z^-$
$E^-T^+Z^-$
Transient
$E^-T^-Z^+$
x 10^{-3}
Predicted vs actual 3-cell correlations
Actual value
Predicted value
Predicted vs actual 3-cell correlations
Log(actual value)
Log(predicted value)

Figure S7


A
0.1
0.1
0.1
0.1
0.1
0.1
0.1
0.1
Frequency
Waiting time (generations)
B
0.1
0.1
0.1
0.1
0.1
Frequency
Waiting time (generations)
C
0.025
0.025
0.025
0.025
0.025
Frequency
Waiting time (generations)

Supplementary Movie 1; related to Figure 2. Time-lapse movie shown in Fig 2B of H2B-mCitrine driven by Esrrb. Closely related cells quickly mix in space with more distantly related cells.

Supplementary Table 1; related to Figures 2-4. Sequences of smFISH probes used to detect transcripts of *Esrrb*, *Tbx3*, and *Zscan4*.

Supplementary Table 2; related to Figures 2 and 3. Sequences of the C-terminus regions of the endogenous genes *Esrrb* and *Tbx3* used as CRISPR/Cas9 targets to make the *Esrrb-mCitrine* reporter line (Figure 2) and the double (*Esrrb-mCherry* and *Tbx3-mCitrine*) reporter line (Figure 3 and S5).

Supplementary Table 3; related to Figure 3. RNA-seq mapped reads for the five subpopulations of cells, Z-, Z+, E-T-, E+T-, and E+T+.

Supplemental Figure Captions

Figure S1. KCA can infer the dynamics of a full range of cell state transition networks; related to Figure 1. **(A)** (i) Chain-like linear transition network. (A.ii) Example pedigree showing typical transitions for the network in A.i in a proliferating colony of cells. (A.iii) Two-cell correlation matrix for sister cells (u=1). (A.iv) Three-cell correlation matrices for triplets of cells at lineage distance u=1,v=2. (A.v) The inferred transition network from the two-cell correlation functions measured over 30 simulated trees of 5 generations (using the framework in Box 1). **(B)** (i) Same as in A but for a transition network comprised of a reversible cycle. (B.ii) A red cell can directly give rise to a progeny in either the blue or green states (red boxes). (Biii-v) same as in A but for the reversible cycle network. **(C)** (i)Three-state cycle with irreversible transition. (C.ii) Only green cells can emerge as progenies of red cells (red boxes). Two-cell correlation functions do not capture the unidirectional nature of these transitions. Note that the two-cell correlation matrices are the

same in B.iii and C.iii. The three-cell correlation functions, however, are different between the two transition networks. In B.iv, the joint probability of observing the more distant relative of a triplet of cells and one of the two more closely related cells in the red state is the same when the third cell is in the green state or the blue state (outlined), consistent with reversible transitions. This is because a red cell can give rise to both green and blue progenies. In C.iv, however, the joint-probability of observing the same configuration is significantly higher when the third cell is in the green state than when it is in the blue state (boxed in red), consistent with the fact that red cells can only give rise to green cells. To infer the dynamics of the irreversible cycle (C.v), we used three-cell correlation functions (see STAR Methods). **(D)** Example of a developmental decision tree with irreversible directional transitions, and non-stationary dynamics. All cells initially start off in the green state and eventually transition to the terminal states at the bottom of the tree. KCA applied to three-cell correlation functions can be used to infer the dynamics of this system, as demonstrated here using 100 simulated trees of size 5 generations (right). **(E)** Recovery of Transition Rates Using KCA from a Cancer System with Irreversible Transitions. Example of measured cell state transition dynamics in cancer cell lines (Gupta et al., 2011). This system violates the condition of reversibility. Deviation of the inferred rates from the actual rates as a function of number of trees simulated. Three-cell correlation functions, but not two-cell correlation functions, are sufficient to infer the full dynamics with arbitrary accuracy, with increasing number of observations. The deviation of the inferred transition matrix from the actual one was computed as $\sqrt{\sum_{ij}\left(T_{ij}^{actual}-T_{ij}^{inferred}\right)^2}$.

Figure S2. Validation of the Esrrb reporter cell line; related to Figure 2.

(A) smFISH analysis of endogenous *Esrrb* transcripts and knock-in Esrrb reporter transcripts in the same cell. Note that transcripts appear in both channels, including both active transcriptional sites (bright foci, blue arrows) and individual cytoplasmic mRNA molecules (dimmer dots). This indicates successful

homozygous targeting of the Esrrb locus. **(B)** *Esrrb* promoter activity (PA) extracted from the final cell cycle of movies before fixation (see Methods) correlates well with *Esrrb* transcript count at final time-point (r=0.71), measured by smFISH (left). Note that β-Actin, a negative control, shows much weaker correlation (r=0.31) with *Esrrb* promoter activity (right). **(C)** In order to sort cells and independently validate the inferred chain-like dynamics with movies, we constructed a second cell line containing non-perturbing reporters for *Esrrb* and *Tbx3* (schematic, see methods for knock-in construction). This line was validated by checking for co-localization of *mCitrine* transcripts with *Tbx3* transcripts (i) and *mCherry* transcripts with *Esrrb* transcripts (ii), and by genomic PCR (not shown). Note that while almost all *mCitrine* transcripts co-localize with *Tbx3* transcripts, indicating a double (homozygous) knock-in, only half the *mCherry* transcripts co-localize with *Esrrb*, indicating that the fluorescent reporter was incorporated into one *Esrrb* allele. **(D)** *Tbx3* promoter activity is correlated with *Tbx3* transcript count. By tracking movies of the *Tbx3*-mCitrine fluorescence, we computed promoter activity and used smFISH to quantify *Tbx3* transcripts at the endpoint. Here, we plot mRNA copy number versus promoter activity averaged over the final cell cycle. Transcript-counts higher than 15 reliably correspond to significant promoter activity.

Figure S3. Robustness of the inference results to the choice of Tbx3 and Zscan4 thresholds; related to Figure 4.

In this figure, we show that the inferred transition rate matrix does not depend sensitively on the choice of transcript-count thresholds used to define the high and low states for the genes *Tbx3* and *Esrrb*. **(A)** We varied the threshold for assigning a cell to the *Tbx3*-high (T+) state from 10 to 30 transcripts. We then repeated the same analysis as in Fig. 4E,F and inferred the transition rate matrix from the observed correlation functions. Each matrix shows the deviation of the inferred transition rate matrix from the reference transition rate matrix inferred using a threshold of 15 transcripts (the threshold used for the

analysis in Fig. 4). The deviations are normalized by the standard deviation of the reference transition rates computed using bootstrap over all the colonies (the error bars shown in Fig. 4Eii,F). The inferred transition rates are not significantly different for *Tbx3* thresholds of 10 to 17 transcripts. For higher thresholds, a significant rate emerges for the transition from the E+T-Z- state to the transient E-T+Z- state. This is because at higher thresholds cells that should be classified as E+T+Z- are instead classified as E+T-Z-. **(B)** The deviation between the inferred transition matrix at a given threshold from the reference matrix (evaluated at threshold 15) computed using the Frobenius distance of the two matrices; Frobenius distance of two matrices A and B is defined as $d(A,B) = \sqrt{\sum_{i,j}\left(A_{i,j} - B_{i,j}\right)^2}$. The dashed line is the deviation expected solely from the statistical uncertainty in the rates of the reference transition matrix that stems from the finite size of the data; two transition matrices inferred using the same thresholds from the same number of observed trees in different iterations of the experiment will on average deviate by this much. At high thresholds, the inferred transition rate matrix is significantly different. To directly verify that a threshold of around 15 indeed corresponds to activation of *Tbx3*, we used a *Tbx3*-mCitrine knock-in cell line (see main text for description and Figure S2D). **(C)** Same analysis as in (A) but with the *Tbx3* threshold fixed at 15 and the *Zscan4* threshold varied from 10 to 250. Because *Zscan4* on cells can produce roughly hundred to thousands of transcripts (Fig. 3A), we do not expect significant sensitivity on the *Zscan4* threshold, as long as the threshold is sufficiently large to capture these events. The plots show the normalized deviation between the transition matrices inferred using various thresholds from the reference matrix that is inferred using a threshold of 50. **(D)** The Frobenius distance between the transition rate matrices at different thresholds from the reference matrix. The inferred transition rates do not change significantly as the threshold is changed from 50 to 250 counts.

Figure S4. Pedigrees reveal limited combinations of states; related to Figure 4.

Pedigrees, gene expression levels, and cell state assignments for the 41 measured trees used to analyze the *Tbx3*, *Esrrb*, and *Zscan4* states. Closely related cells are typically found in the same state. Specific combinations of states occur frequently (e.g. E-T-Z- (blue) cells are related to E+T-Z- (red) cells). The gene expression columns correspond to expression levels of *Esrrb*, *Tbx3*, and *Zscan4* from left to right.

Figure S5. Inference and validation of a five-state network in ESCs; related to Figure 4.

(A) The two-cell correlation matrix of the trees in Figure S4 for pairs of cells at lineage distance u = 2 (top) and lineage distance u = 3 (bottom). **(B)** For each of the five states in Figure 3, we plot the inferred state stability (diagonal terms of the transition matrix) as a function of lineage distance, u. Dashed line serves as a guide for the eye to indicate constant rates over time. Note that rates are constant with respect to u within error bars, estimated by bootstrap. **(C)** To verify that the chain-like model (Fig. 4F) is consistent with the two-state approximation of *Esrrb* dynamics in Figure 2, we simulated the dynamics of the 5-state chain for 14 trees of 4 generations each. We discarded *Tbx3* and *Zscan4* information, and assigned each cell to a state only based on its *Esrrb* level. KCA was used to infer the transition rates between the two composite *Esrrb* states. Simulations were repeated 10,000 times to estimate the statistical error of the inferred rates. Note that inferred rates were consistent with measured rates for *Esrrb* dynamics (Fig. 2Eiv) (blue line indicates low to high transition rate, and red line indicates high to low transition rate). Thus, although *Esrrb* dynamics are not precisely described by a two-state Markovian process, they are nevertheless well-approximated by such a model within the statistical limits of the data sets obtained here. At the same time, the 5-state linear chain inferred in Figure 4 provides a more accurate description of these dynamics (see STAR Methods). **(D)** The inferred 5-state chain (Fig. 4F) predicts that E-T-Z+ cells emerge from E+T+Z- via a transient state (black circle). As a result, *Esrrb* and *Tbx3* should both turn off almost simultaneously just prior to *Zscan4* activation. Using the *Tbx3*/Esrrb double reporter line

(Fig. S2C), we observed these predicted dynamics in movies. Here, three example of abrupt and nearly simultaneous shut-off of both *Esrrb* and *Tbx3* near the end of the movie (event indicated by the arrowheads, and subsequent plateauing of total fluorescence in each trace, consistent with no further fluorescent protein expression), and subsequent expression of *Zscan4* by smFISH (right); also shown is the DAPI stained nucleus of the same cells.

Figure S6. Three-point analysis of the ES cell state transition network; related to Figure 4.

Comparison of the two-cell and the three-cell correlation analysis. The predicted and the observed three-cell correlation functions. Left column: the calculated three-cell correlation functions for a triplet of cells at lineage distances u=1 and v=2, computed from the transition rates inferred from the two-cell correlation functions shown in Figure 4E. See STAR Methods for how the three-cell correlation functions are calculated from the transition matrix. The three-cell correlations are 5x5x5 matrices, represented here as 5 square 5x5 matrices. The second column shows the actual three-cell correlations measured directly using the observed trees (Figure S4). The third column shows the corresponding statistical error (fractional error of each entry) in the rightmost column. Notice that the predicted three-cell correlation functions are consistent with the observed three-cell correlation functions, validating the inferred chain. This is also shown in the plots at the bottom. Left plot shows the predicted versus actual three-cell correlations, and statistical errors. In this plot, each point represents one of the 75 independent matrix elements (for N states, there are N^2(N+1)/2 unique matrix elements). (Right) Plot of the logarithm (base 10) of the predicted and actual three-cell correlations. Only values larger than 0.001 are plotted. Note also that there are no free parameters in this figure. The agreement between the predicted and actual three-cell correlations is obtained without any fitting.

***Figure S7. Distribution of waiting times between consecutive visits to a given state varies between different network motifs; related to Figure 4 and STAR Methods section "*Potential benefits of chain-like state transition networks."**

The dynamics of the three state transition networks shown here, A) Chain-like linear B) Irreversible cycle, and C) All-to-all, were simulated. Every time a simulated cell visited the green state, we tabulated the number of generations that had elapsed since its most recent ancestor left that state. From this, we computed the distribution of "waiting times" between consecutive visits to the green state for each type of network (blue bar plot). As expected, the mean waiting time (blue line) is the same for all three networks. The transition rates were selected to ensure that the flux of cells into the green state and the population fractions are the same across the networks, also ensuring identical mean waiting-times. The distribution of waiting-times, however, differs significantly between the networks. In particular, the chain-like linear network exhibits the smallest median (red line), implying that most cells in the population have visited the green states relatively recently, at the expense of a minority that have experienced larger than average waiting-times. Assuming that visiting the green state enhances the viability of the cells, this could be beneficial for the culture as a whole.